PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-248494

(43)Date of publication of application : 05.09.2003

(51)Int.Cl. G10L 11/00
G06F 17/30
G10L 15/00
G10L 15/02
G10L 15/10

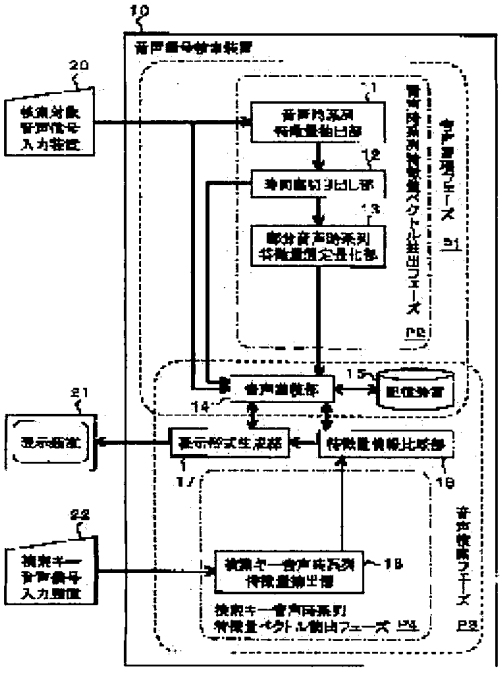
(21)Application number : 2002-047806

(22)Date of filing : 25.02.2002

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

(72)Inventor : SUGA HIROTOSHI
TERAMOTO JUNJI
KATAOKA RYOJI

(54) AURAL SIGNAL RETRIEVING METHOD, AURAL SIGNAL STORING METHOD FOR SPEECH RETRIEVAL, AURAL SIGNAL RETRIEVING DEVICE, PROGRAM THEREFOR AND RECORDING MEDIUM FOR THE PROGRAM



(57)Abstract:

PROBLEM TO BE SOLVED: To make retrievable speech data similar to a retrieval key at a high speed by shortening the time of calculation of a similar distance by a simple matching between fixed-length data in retrieval of the speech data.

SOLUTION: A time window segmentation part 12 extracts partial speech time-series feature quantities which are different in length from an aural signal to be retrieved by using time windows of a plurality of kinds of lengths and a partial speech time- series feature quantity length fixation part 13 linearly expands or contracts partial speech time-series feature quantities to a specified reference time window length and stores them in a speech storage part 14. In retrieval, a retrieval key speech time-series feature quantity extraction part 18 extracts a retrieval key speech time-series feature quantity vector with the length of a reference time window from the input speech signal of the retrieval key and a feature quantity information comparison part 16

calculates the similarity distance between the extracted vector and a speech time-series feature quantity vector to be retrieved to decide a vector having higher similarity as a retrieval result.

* NOTICES *

CLAIMS

[Claim(s)]

[Claim 1]A process in which partial tone voice time series characteristic quantity from which length differs using a time window of two or more kinds of length, respectively is extracted from an audio signal used as a retrieval object, Linearity elasticity of the partial tone voice time series characteristic quantity of two or more kinds of extracted length is carried out, A process arranged with the length of a base period window used as a standard of similar distance calculation at the time of search, A process accumulated as a voice time series feature amount vector used as a retrieval object for measuring partial tone voice time series characteristic quantity arranged with the length of said base period window with a voice time series feature amount vector obtained from an audio signal into which it is inputted as a search key, A process in which a search key voice time series feature amount vector of the length of said base period window is extracted from an audio signal inputted or specified as a search key, A process in which calculate similar distance of a voice time series feature amount vector accumulated as said retrieval object, and said search key voice time series feature amount vector, and similarity of an audio signal of a search key and an audio signal of each audio signal section which is a retrieval object is computed, An audio signal search method having a process in which search results are outputted based on a computed result of similarity.

[Claim 2]A process in which partial tone voice time series characteristic quantity is extracted from an audio signal used as a retrieval object using a base period window of length used as a standard of similar distance calculation at the time of search, A process accumulated as a voice time series feature amount vector used as a retrieval object for measuring extracted partial tone voice time series characteristic quantity with a voice time series feature amount vector obtained from an audio signal into which it is inputted as a search key, A process in which partial tone voice time series characteristic quantity from which length differs using a time window of two or more kinds of length, respectively is extracted from an audio signal inputted or specified as a search key, Linearity elasticity of the partial tone voice time series characteristic quantity of two or more kinds of extracted length is carried out, Make into a search key voice time series feature amount vector partial tone voice time series characteristic quantity arranged with a process arranged with

the length of said base period window, and the length of said base period window, and as said retrieval object. Similar distance of an accumulated voice time series feature amount vector and said search key voice time series feature amount vector is calculated, An audio signal search method having a process in which similarity of an audio signal of a search key and an audio signal of each audio signal section which is a retrieval object is computed, and a process in which search results are outputted based on a computed result of similarity.

[Claim 3]In a process in which partial tone voice time series characteristic quantity from which length differs using a time window of two or more kinds of length, respectively is extracted from said audio signal. Shifting a base period window of length which extracts voice time series characteristic quantity from said audio signal, and serves as said standard from there little by little. The audio signal search method according to claim 1 or 2 starting partial tone voice time series characteristic quantity, and starting partial tone voice time series characteristic quantity shifting little by little similarly further by a time window of two or more kinds of length centering on the length of a base period window.

[Claim 4]Claim 1, wherein a voice time series feature amount vector accumulated as said retrieval object and said search key voice time series feature amount vector carry out linearity compression of the partial tone voice time series characteristic quantity of the length of a base period window at predetermined length, the audio signal search method according to claim 2 or 3.

[Claim 5]Input information which specifies the outputted audio signal section as search results as said search key once searching, and a voice time series feature amount vector corresponding to the specified audio signal section is made into said search key voice time series feature amount vector, An audio signal search method given in either from claim 1 performing re retrieval by similar distance calculation with a voice time series feature amount vector accumulated as said retrieval object to claim 4.

[Claim 6]A process in which an audio signal used as a retrieval object is inputted, and a process in which voice time series characteristic quantity is extracted from an inputted audio signal, A process in which partial tone voice time series characteristic quantity is started while shifting a base period window which is a time window of length which serves as a standard from said voice time series characteristic quantity little by little, A process in which partial tone voice time series characteristic quantity is started while shifting little by little also about a time window of two or more kinds of length centering on the length of said base period window, A process which carries out linearity elasticity of the partial tone voice time series characteristic quantity of said two or more kinds of length, and is arranged with the length of a base period window, Partial tone voice time series characteristic quantity arranged with the length of said base period window as a search key. An audio signal accumulating method for voice search having a process accumulated as a voice time series feature amount vector used as a retrieval object for comparing with a voice time series feature amount vector obtained from an audio signal inputted.

[Claim 7]An audio signal accumulating method for the voice search according to claim 6 making what carried out linearity compression of the partial tone voice time series characteristic quantity arranged with the length of said base period window at predetermined length into a voice time series feature amount vector of a retrieval object to accumulate.

[Claim 8]A means to extract partial tone voice time series characteristic quantity from which length differs using a time window of two or more kinds of length, respectively from an audio signal used as a retrieval

object, Linearity elasticity of the partial tone voice time series characteristic quantity of two or more kinds of extracted length is carried out, A means arranged with the length of a base period window used as a standard of similar distance calculation at the time of search, A means to accumulate as a voice time series feature amount vector used as a retrieval object for measuring partial tone voice time series characteristic quantity arranged with the length of said base period window with a voice time series feature amount vector obtained from an audio signal into which it is inputted as a search key, A means to extract a search key voice time series feature amount vector of the length of said base period window from an audio signal inputted or specified as a search key, A means to calculate similar distance of a voice time series feature amount vector accumulated as said retrieval object, and said search key voice time series feature amount vector, and to compute similarity of an audio signal of a search key, and an audio signal of each audio signal section which is a retrieval object, An audio signal retrieval device provided with a means to output search results based on a computed result of similarity.

[Claim 9]A means to extract partial tone voice time series characteristic quantity from an audio signal used as a retrieval object using a base period window of length used as a standard of similar distance calculation at the time of search, A means to accumulate as a voice time series feature amount vector used as a retrieval object for measuring extracted partial tone voice time series characteristic quantity with a voice time series feature amount vector obtained from an audio signal into which it is inputted as a search key, A means to extract partial tone voice time series characteristic quantity from which length differs using a time window of two or more kinds of length, respectively from an audio signal inputted or specified as a search key, Linearity elasticity of the partial tone voice time series characteristic quantity of two or more kinds of extracted length is carried out, A means arranged with the length of a base period window used as a standard of similar distance calculation at the time of search, Make into a search key voice time series feature amount vector partial tone voice time series characteristic quantity arranged with the length of said base period window, and as said retrieval object. Similar distance of an accumulated voice time series feature amount vector and said search key voice time series feature amount vector is calculated, An audio signal retrieval device provided with a means to compute similarity of an audio signal of a search key, and an audio signal of each audio signal section which is a retrieval object, and a means to output search results based on a computed result of similarity.

[Claim 10]A program for audio signal search for making either from claim 1 to claim 5 perform an audio signal search method of a statement to a computer.

[Claim 11]A recording medium of a program for audio signal search recording a program for making either from claim 1 to claim 5 perform an audio signal search method of a statement to a computer.

[Claim 12]A program for audio signal accumulation for making a computer perform an audio signal accumulating method for the voice search according to claim 6 or 7.

[Claim 13]A recording medium of a program for audio signal accumulation recording a program for making a computer perform an audio signal accumulating method for the voice search according to claim 6 or 7.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001]

[Field of the Invention]Especially this invention about the art of a voice search system An input. Or it is related with the audio signal accumulating method for the audio signal search method for searching the audio signal section similar to the specified audio signal out of a voice database, and voice search, an audio signal retrieval device, its program, and the recording medium of the program.

[0002]

[Description of the Prior Art]As conventional technology of the technique of searching an audio signal similar to an input voice signal out of a voice database, there are some which were stated to the following reference 1.

[Reference 1] Takashi Endo, Masayuki Nakazawa, Hironobu Takahashi, the Japanese Society for Artificial Intelligence national conference (the 12th time) in the sound, data representation [by the self-organization network of video], and ******:"both spotting search": 1998 fiscal year, S5-04, pp.122-125.

This is IPM (Incremental Path Method). It is the method of searching the audio signal which calculated the similarity between audio signals by DP matching and the same dynamic matching, and was similar to the input voice signal using the network out of a voice database. According to this method, similarity is calculable to a time base direction by using dynamic matching also with the audio signal which is carrying out nonlinear elasticity.

[0003]

[Problem(s) to be Solved by the Invention]When it is going to search an audio signal with an audio signal, it is necessary to calculate the similarity between audio signals. Generally, even when the same character string is uttered, nonlinear elasticity of the audio signal is carried out on a time-axis. In order to calculate the similarity between audio signals and to correspond to nonlinear elasticity from this, dynamic matching of DP matching, Hidden Markov Model (HMM), etc. needed to be used.

[0004]However, all audio signals do not carry out nonlinear elasticity on a time-axis. As an example of the audio signal which does not carry out nonlinear elasticity, singing voice is mentioned first. although the tempo of the whole section may come to be alike slowly or may become quick if it is the short-time section since tempo exists in singing voice potentially, tempo is not confused within the section. That is, if singing voice is within the short-time section, although it may carry out linearity elasticity in the whole section, nonlinear elasticity of it is not carried out within the section.

[0005]There is announcer's utterance speech as other examples of the audio signal which does not carry out nonlinear elasticity. Since announcer is very good at uttering the same language at the same tune repeatedly, if he is the same language, he will not do nonlinear elasticity. If ordinary persons' utterance is also short utterance of about one word, even if it may carry out linearity elasticity in the whole section, within the section, it can be considered that nonlinear elasticity has hardly been carried out.

[0006]In the Prior art, when carrying out similarity calculation between audio signals, dynamic matching was used. Therefore, dynamic matching which can respond to the sound which has carried out only linearity elasticity, and nonlinear elasticity will be performed. There was a problem that dynamic matching will have much computational complexity compared with static matching which performs distance calculation, such as Euclidean distance between fixed-length vectors and a Manhattan distance, and search time will become

long.

[0007]

[Means for Solving the Problem]In order to solve said technical problem, this invention enables it to compare characteristic quantity started using a time window in search of voice data at high speed not by dynamic matching but by static matching. Therefore, linearity elasticity of the voice data of voice time series characteristic quantity is carried out, and step is kept with fixed length data of the length of a certain fixed time intervals. By carrying out like this, search of a search key and similar voice data is enabled at high speed by simple matching between fixed length data.

[0008]Specifically, an audio signal retrieval device of this invention consists of a voice time series feature amount extracting means, a time window logging means, a partial tone voice time series characteristic quantity fixed length-ized means, a search key voice time series feature amount extracting means, a voice storage means, a search condition input means, a characteristic quantity information comparison means, and a display style creating means. A voice time series characteristic quantity linearity compression means and a search key voice time series characteristic quantity linearity compression means may be established.

[0009]A voice time series feature amount extracting means extracts voice time series characteristic quantity from an audio signal which is a time series signal.

[0010]A time window logging means prepares a base period window which is a time window of length which serves as a standard first. And partial tone voice time series characteristic quantity of the length of a base period window is started from voice time series characteristic quantity extracted by a voice time series feature amount extracting means, shifting a base period window little by little. Next, a time window of two or more kinds of length is prepared focusing on the length of a base period window. Partial tone voice time series characteristic quantity is similarly started by a time window of such length.

[0011]A partial tone voice time series characteristic quantity fixed length-ized means generates partial tone voice time series characteristic quantity which carried out linearity elasticity and arranged with base period window length partial tone voice time series characteristic quantity started by a time window of two or more kinds of length by a time window logging means, and extracts it as a voice time series feature amount vector.

[0012]A voice time series characteristic quantity linearity compression means carries out linearity compression of the voice time series characteristic quantity of base period window length generated by a partial tone voice time series characteristic quantity fixed length-ized means at a certain fixed length, and extracts a voice time series feature amount vector.

[0013]A search key voice time series feature amount extracting means extracts search key voice system sequence characteristic quantity of base period window length, and extracts it from an audio signal of the length of a base period window inputted as a search key as a search key voice time series feature amount vector.

[0014]A search key voice time series characteristic quantity linearity compression means carries out linearity compression of the voice time series characteristic quantity of base period window length extracted by a search key voice time series feature amount extracting means at the same length as a voice time series feature amount vector, and extracts a search key voice time series feature amount vector.

[0015]From a voice time series feature amount vector which a voice storage means accumulated an

inputted audio signal of a retrieval object, and was extracted by a voice time series characteristic quantity linearity compression means, create an index and with a voice time series feature amount vector. It accumulates and time window logging information which is information on matching with the audio signal section of extraction origin of a voice time series feature amount vector further is also accumulated.
[0016]When a search condition input means uses as a search key the audio signal section accumulated in a voice storage means, it inputs conditions for specifying a search key.
[0017]A characteristic quantity information comparison means calculates similar distance for a search key voice time series feature amount vector extracted in a search key voice time series characteristic quantity linearity compression means, and a voice time series feature amount vector accumulated in a voice storage means by static matching, and sets up similarity. Thereby, similarity calculation more nearly high-speed than dynamic matching becomes possible. And similarity with a search key voice time series feature amount vector outputs a voice time series voice feature amount vector to high order.
[0018]A display style creating means matches a voice time series feature amount vector which at least order was attached and was outputted from a characteristic quantity information comparison means with the audio signal section of extraction-time window logging information accumulated in voice storage means origin to origin, and outputs it to a display.
[0019]
[Embodiment of the Invention][Embodiment 1] Drawing 1 is a lineblock diagram for *describing* the embodiment of the invention 1. According to Embodiment 1, the sound inputted from the outside is used as a search key.
[0020]Operation of this invention comprises the voice storage phase P1, the voice time series feature amount vector extraction phase P2 called from it, the voice search phase P3, and the search key voice time series feature amount vector extraction phase P4 called from it. Hereafter, operation of each phase is explained.
[0021][A]The voice storage phase P1 and voice time series feature amount vector extraction phase P2 drawing 2 are the flow charts explaining operation of the voice storage phase P1 and the voice time series feature amount vector extraction phase P2.
[0022]First, the audio signal of a retrieval object is inputted from the retrieval object sound signal input device 20, and the audio signal of a voice storage part 14 smell lever is accumulated in the memory storage 15 (Step S1).
[0023]Next, in the voice time series feature quantity extracting part 11, voice time series characteristic quantity is extracted from the inputted audio signal (Step S2). As voice time series characteristic quantity, the speech power of the low following paragraph of the Meru frequency cepstrum coefficient, its primary difference, and each [according to difference, speech power and filter bank analysis the 2nd order] zone, etc. can be expressed with a multi dimensional vector, and what arranged them in order of the time series can be used, for example. The example of voice time series characteristic quantity is stated to the following reference 2.
[Reference 2] : "IT text voice recognition system" besides Kiyohiro Kano, Ohm-Sha, 2001.
Next, in the time window logging part 12, the partial tone voice time series characteristic quantity of base period window length is started, setting up the time window of the length used as a standard as a base

period window, and shifting this base period window little by little, as first shown in drawing 3 (Step S3).
[0024]As shown in drawing 4, the time window of two or more kinds of length centering on the length of a base period window is set up, and the time window length's partial tone voice time series characteristic quantity is started like the case of a base period window, respectively, shifting a time window little by little (step S4). In the example of an experiment mentioned later, the number of base period windows is 150, and they are about 26 milliseconds per frame in length. The minimum of the length of a time window considered it as 118 frames, and the maximum was made into 182 frames. If partial tone voice time series characteristic quantity is started using these time windows, the information about logging of the time window will be accumulated in the memory storage 15 as time window logging information in the voice storage part 14.
[0025]In the partial tone voice time series characteristic quantity fixed length-ized part 13, as shown in drawing 5, Linearity elasticity of the partial tone voice time series characteristic quantity started by the time window of two or more kinds of length is carried out in a time base direction, respectively, step is kept with the length of a base period window (Step S5), and the partial tone voice time series characteristic quantity of the base period window length is generated, and let it be a voice time series feature amount vector (Step S6).
[0026]And an index is built from the obtained voice time series feature amount vector, and it accumulates in the memory storage 15 of the voice storage part 14 with a voice time series feature amount vector (Step S7). As an index structure of the multi-dimension spatial vector, SR-tree stated to the following reference 3, A-tree stated to the reference 4, etc. can be used, for example.
[Reference 3] Norio Katayama. and Shin'ichi Satoh:"The. SR-Tree :. [An Index Structure] for High-Dimensional. Nearest Neighbor Queries",In Proc. ACM SIGMOID International Conference on Management of Data ,pp.368-380,May 1997.
[Reference 4] Yasushi Sakurai, Masatoshi Yoshikawa, Shunsuke Uemura, and Haruhiko Kojima : "The A-Tree:An Index. Structure for High-Dimensional Spaces UsingRelative Approximation" and In Proc. of the 26th International. Conference on Very Large Data Bases(VLDB),pp.516-526,Cairo ,September 2000.
[B]The voice search phase P3 and search key voice time series feature amount vector extraction phase P4 drawing 6 are the flow charts explaining operation of the voice search phase P3 and the search key voice time series feature amount vector extraction phase P4.
[0027]First, the audio signal of the base period window length which becomes a search key is inputted using the search key sound signal input device 22 (Step S10).
[0028]Next, in the search key voice time series feature quantity extracting part 18, the search key voice time series characteristic quantity of base period window length is extracted from the audio signal of the base period window length which inputted (Step S11). The characteristic quantity same as search key voice time series characteristic quantity as the voice time series characteristic quantity of the voice storage phase P1 mentioned above is used. Let this search key voice time series characteristic quantity be a search key voice time series feature amount vector (Step S12).
[0029]In the characteristic quantity information comparing element 16, the similar distance of the obtained search key voice time series feature amount vector and the voice time series feature amount vector accumulated in the memory storage 15 in the voice storage part 14 is calculated (Step S13). It carries out to this distance calculation using the index accumulated in the memory storage 15 of the voice storage part

14 using static matching of the Euclidean distance between fixed-length vectors, a Manhattan distance, etc. Thereby, distance calculation more nearly high-speed than dynamic matching becomes possible. And at least order attaches the audio signal section of a voice time series feature amount vector to order with the short distance (Step S14).

[0030]Finally, in the display style generation part 17, a voice time series feature amount vector is matched with the audio signal section of the extraction origin using the time window logging information accumulated in the memory storage 15 by the voice storage part 14, and it outputs to the display 21 (Step S15). The list information which shows the how many seconds it is from the head of a program, and the audio signal section in an order from the high result of ranking when displaying the search results of the portion which corresponds to a search key, for example out of the musical program of 1 hour in a display here, When the button for reproduction of the portion is displayed and a reproduction button is pushed, it performs outputting the sound of the portion. By this, a retrieving person can save now the time and effort which discovers the audio signal section which suits a retrieval object. When it has information, including a track name etc., in a database about a retrieval object, the track name of search results, etc. can be displayed collectively.

[0031][Embodiment 2] Drawing 7 is a lineblock diagram for describing the embodiment of the invention 2. Embodiment 2 makes it possible to compress the characteristic quantity started using the time window, to make size small, and to search by building the smaller database of size.

[0032]What carried out linearity compression of the partial tone voice time series characteristic quantity which carried out linearity elasticity, and which was arranged with the length of the base period window in Embodiment 2 on the time-axis is made into a voice time series feature amount vector. It uses and what carried out linearity compression of the search key voice time series characteristic quantity of the length of a base period window on the time-axis at the same length as a voice time series feature amount vector is used as a search key voice time series feature amount vector in connection with this. Therefore, compared with Embodiment 1, the voice time series feature amount vector extraction phase P2 called from the voice storage phase P1 and the search key voice time series feature amount vector extraction phase P4 called from the voice search phase P3 differ, and are as follows.

[0033][A]The voice storage phase P1 and voice time series feature amount vector extraction phase P2 drawing 8 are the flow charts explaining operation of the voice storage phase P1 and the voice time series feature amount vector extraction phase P2.

[0034]An audio signal is inputted from the retrieval object sound signal input device 20 like Embodiment 1 (Step S20). In the voice time series feature quantity extracting part 11, voice time series characteristic quantity is extracted from the inputted audio signal (Step S21), and in the time window logging part 12, the partial tone voice time series characteristic quantity of base period window length is started, shifting a base period window little by little (Step S22). In [start partial tone voice time series characteristic quantity also by the time window of two or more kinds of length (Step S23), and] the partial tone voice time series feature fixed length-ized part 13, Linearity elasticity of the partial tone voice time series characteristic quantity of two or more kinds of started length is carried out on a time-axis, and the partial tone voice time series characteristic quantity arranged with base period window length is generated (Step S24).

[0035]Next, in the voice time series characteristic quantity linearity compression zone 30, as shown in

drawing 9, linearity compression of the partial tone voice time series characteristic quantity arranged with the length of the base period window is carried out in a time base direction, respectively, and a voice time series feature amount vector is extracted (Step S25). By this, the number of dimension of a voice time series feature amount vector can become small, the computational complexity of similarity calculation can be reduced, and capacity of the memory storage 15 to accumulate can also be made small.

[0036]An index is built from the voice time series feature amount vector produced by making it be the same as that of Embodiment 1, and it accumulates in the memory storage 15 of the voice storage part 14 with a voice time series feature amount vector.

[0037][B]The voice search phase P3 and search key voice time series feature amount vector extraction phase P4 drawing 10 are the flow charts explaining operation of the voice search phase P3 and the search key voice time series feature amount vector extraction phase P4.

[0038]In [input the audio signal of the base period window length which becomes a search key like Embodiment 1 using the search key sound signal input device 22 (Step S30), and] the search key voice time series feature quantity extracting part 18, The search key voice time series characteristic quantity of base period window length is extracted from the audio signal of the base period window length which inputted (Step S31).

[0039]And in the search key voice time series characteristic quantity linearity compression zone 31, as shown in drawing 9, linearity compression of the voice time series characteristic quantity of base period window length is carried out in a time base direction, and a search key voice time series feature amount vector is extracted (Step S32). Let the length of a search key voice time series feature amount vector be the same length as the voice time series feature amount vector generated in said voice storage phase P1.

[0040]The search key voice time series feature amount vector obtained in the characteristic quantity information comparing element 16 like Embodiment 1, Similar distance with the voice time series feature amount vector accumulated in the memory storage 15 in the voice storage part 14 is calculated (Step S33), and at least order attaches a voice time series feature amount vector to order with the short distance (Step S34).

[0041]Finally, like Embodiment 1, in the display style generation part 17, a voice time series feature amount vector is matched with the audio signal section of the extraction origin using the time window logging information accumulated in the memory storage 15 of the voice storage part 14, and it outputs to the display 21 (Step S35). For example, the list information which shows the how many seconds it is from the head of a program, and the audio signal section in an order from the high result of ranking when displaying the search results of the portion which corresponds to a search key out of the musical program of 1 hour, The button for reproduction of the portion is displayed, and when a reproduction button is pushed, it displays by the display style which can output the sound of the portion.

[0042][Embodiment 3] Drawing 11 and drawing 12 are the lineblock diagrams for describing the embodiment of the invention 3. Embodiment 3 does not input a search key from the outside, but when it searches once and search results are obtained, it newly respecifies a search key out of search results, and it enables it to search voice data similar to it.

[0043]In order of similarity, two or more search results which use voice data as a search key are displayed on a screen, as shown in a table, and other similar voice data is searched by using as a new search key voice

data obtained as search results.

[0044]For example, if the portion of the "track name" of a search-results display screen is touched in music search, the musical piece will be chosen as search results, and voice response will be played and carried out from the head of the portion corresponding to a search key, or a musical piece. If the portion of the "ranking" on a screen is touched, what performs a search of a further similar musical piece by using as a new search key the search results (for example, the voice data which is the original search key among musical pieces and a similar portion (for 3 to 4 seconds)) of the musical piece will be performed.

[0045]As mentioned above, the result that it was the closest to the retrieval object is chosen as a search key from search results, and it refers to Embodiment 3 again. The voice storage phase P1 of the audio signal retrieval device 10 shown in drawing 11 is the same as the voice storage phase P1 of Embodiment 1 shown in drawing 1, and the voice storage phase P1 of the audio signal retrieval device 10 shown in drawing 12 of it is the same as the voice storage phase P1 of Embodiment 2 shown in drawing 7. In Embodiment 1 mentioned above and Embodiment 2, the voice search phase P3 differs and are as follows.

[0046][A]Voice search phase P3 drawing 13 is a flow chart explaining operation of the voice search phase P3.

[0047]As a preceding paragraph story, it searches like Embodiment 1 and Embodiment 2, at least order attaches the search results in order of similarity, and it displays on the display 21 (Step S40). Search will be ended if the displayed search results have fully agreed in the retrieval object with the directions from a user (Step S41).

[0048]If it has not agreed in a retrieval object, in the search condition input section 23, the result nearest to the retrieval object of the search results to which at least the order already displayed was attached is made to specify as a search key, and it is re(Step S42) chosen as a search key. For this reason, when displaying the result to which at least order was attached in Step S40 with the display 21, two buttons for an input are displayed per result. One is a button pushed when specifying as a search key, and one is a button pushed when making the sound of the result utter. The user can specify a search key out of search results by pushing the former button on the display 21.

[0049]If a search key is specified, in the characteristic quantity information comparing element 16, the accumulated high voice interval of a search key and similarity will be searched, and ranking will be carried out to the high order of similarity (Step S43). In the display style generation part 17, the display style which can choose a search key from search results is generated, and it displays on the display 21 (Step S40).

[0050]If the displayed search results have fully agreed in the retrieval object, they will end search (Step S41). In fully not agreeing yet, a search key is rechosen further and it searches again, and it repeats until the result which fully agrees in a retrieval object is obtained (Steps S40-S43).

[0051][Embodiment 4] In above Embodiments 1-3, extract the partial tone voice time series characteristic quantity from which length differs using the time window of two or more kinds of length, respectively about the audio signal of a retrieval object, and linearity elasticity of those partial tone voice time series characteristic quantity is carried out. What was arranged with the length of the base period window was accumulated by the voice storage part 14 as a voice time series feature amount vector of a retrieval object.

[0052]According to Embodiment 4, linearity elasticity of the partial tone voice time series characteristic quantity started by the time window is not performed about the thing of a retrieval object, but is performed

about what was inputted as a search key. That is, in Embodiment 4, partial tone voice time series characteristic quantity is not extracted from the audio signal of a retrieval object using the time window of two or more kinds of length, but partial tone voice time series characteristic quantity is extracted only using a base period window, and it is accumulated in the voice storage part 14 as a voice time series feature amount vector of a retrieval object. On the other hand, about the audio signal inputted as a search key, the partial tone voice time series characteristic quantity from which length differs using the time window of two or more kinds of length, respectively is extracted, and linearity elasticity of the partial tone voice time series characteristic quantity from which such length differs is carried out so that it may become the length of a base period window.

[0053]Even if it follows that into which linearity elasticity was inputted as a search key, the same search results as Embodiments 1-3 are obtained. After keeping step with the length of a base period window, reduction of the data volume compared with fixed length by carrying out linearity compression like Embodiment 2 at the time of search can also be aimed at if needed.

[0054][Experimental result] In order to check the validity of this invention, two kinds of experiments for the singing data created to the experiment about Embodiment 2 were conducted.

[0055]First, in the 1st experiment, compared with matching which does not correspond to linearity elasticity, in search of singing voice, matching corresponding to the linearity elasticity by this invention checks an effective thing, and then in the 2nd experiment. Even if compared with the conventional system which mounts nonlinear elastic matching, it checked that this invention was sufficiently effective in search of singing voice.

[0056][A]The test subject of 15 experimental conditions common to both experiments is divided into a woman, a male, and three mixed groups, and a 62 song name list is passed to each group. In a song name list, I get those who know the song to sing a part of phrase (about 10 seconds), and 62x3 singing voice (a total of about 30 minutes) is stored in a database.

[0057]About (base-period window length: 150-frame about 4 seconds) one phrase is arbitrarily taken out from 12 music selection and there out of one test subject group's singing voice, and it is considered as a search key. And the same phrase portion of the singing voice other two groups' test subject is searched as conformity results.

[0058]Suppose that singing voice is searched from the voice data of 44100 Hz of sampling frequencies, the quantifying bit number of 16 bits, and the wave file format of one channel in this experiment. As a voice feature amount, the 5-dimensional low following paragraph of the Meru frequency cepstrum coefficient is used, and time series characteristic quantity is extracted from what arranged this in on a time-axis using a time window.

[0059]An average of average search length is used as a valuation basis of search results. This is a valuation basis showing the time and effort which discovers the result which suits a retrieval object out of search results. It should search, even if we decided to judge the conformity up to the 20th place among the search results by which ranking was carried out, ranking was carried out to to the 20th place and less than it had conformity results.

[0060][B]explanation of an average of average search length -- here, an average of the average search length which used as a valuation basis of search results is explained. Average search length is stated to the

following reference 5.

[Reference 5] Tokunaga ****: "language, calculation 5 information retrieval, and language processing", the University of Tokyo publication, 1999.

Average search length is a measure which evaluates the set to which at least order was attached as search results. When the result to which at least order was attached as search results has returned, the retrieving person has to judge the conformity of search results in detail from the result of a higher rank actually. Average search length is a measure which measures the time and effort of the user which must judge the conformity of a result, in order for a retrieving person to obtain a required number of conformity results in consideration of the process of such a retrieving person's conformity judgment.

[0061]For example, suppose that search results were able to divide into set S_1 with an order, S_2, and S_3 as follows. However, an order during a set is the order of S_1, S_2, and S_3, and O and x express conformity results and an incongruent result, respectively.

[0062]S_1: {O, x, x, x}

S_2: {O, O, O, O, x, x}

S_3: {O, O, x, x}

Now, suppose that a retrieving person wants to obtain one conformity results. First, set S_1 will be inspected. Since an order is not attached in this set, the expected value of the number of the result which must be inspected by the time it finds conformity results is set to 4+2x1 [1x1/]/4+3x1/4+4x1/4=2.5.

[0063]In order for this to find one conformity results from search results, the retrieving person shows that the conformity of 2.5 search results must be judged on the average. That is, the number of average search length required to find out one conformity results from these search results is 2.5.

[0064]Since what is necessary is just to find one from set S_2 after all inspecting set S_1 in order to find two conformity results, the expected value of the number of the result which should be inspected is set to x(4+1)4/6+(4+2) x4/15+(4+3) x1/15=5.4. That is, the number of average search length required to find out two conformity results is 5.4.

[0065]In the above-mentioned example, although it was a case where search results were given by the set to which at least order was attached, even when the total order is attached to each of search results, if the element of each set is considered to be one, average search length can be calculated.

[0066]Average search length does not become one measure, but becomes a value depending on the number of required conformity results so that the above thing may also show. Then, the value of the average search length per required conformity results is calculated as an average of average search length.

[0067]When average search length required to find [the required number of conformity results] M and i required conformity results for i and the total number of conformity results is made into x(i), an average of x_{av} of average search length is $x_{av}=(1/M)\ \mathrm{sigma}_{i=1}^{M}\ \{x(i)/i\}$.

It is come out and expressed.

[0068]For example, the case where the result of having suited is searched by the 2nd place and the 6th place is considered as a result of search. When the number of required conformity results is set to 1, average search length is set to 2, and when the number of required search results is set to 2, average search length is set to 6. An average of these average search length is set to (2/1+6/2) / 2=4.

[0069][C]It experiments by the method of extracting time series characteristic quantity as a method which

does not correspond to the 1st experiment linearity elasticity only using the time window of one kind of length. On the other hand, it experiments as a method corresponding to linearity elasticity by the method by this invention using nine kinds of time windows. The value of an average of this average search length that is both a method is compared.

[0070][D]The result of having compared the average of the average search length by the 1st experiment as a result of the 1st experiment is shown in drawing 14. x in a figure means that conformity results were not able to be searched. Although an average of average search length becomes the same value in the song B, E, and H, in the other song, the method by this invention has exceeded it altogether. That is, when the case where the linearity elasticity by the conventional method was not used for the voice data of a retrieval object by a fixed-length time window was compared with the case where the linearity elasticity by this invention is used, it turned out that average search length is short, and it ends, and is effective in retrieval precision improving by about 2 times from 25%. Therefore, the method corresponding to the linearity elasticity by this invention can be said to be effective in search of singing voice.

[0071][E]An average of the average search length of the conventional method and the method by this invention is compared using the voice search service of "CrossMediator for VideoV2.0 (R1)" of media drive incorporated company as the conventional method which can respond to the 2nd experiment nonlinear elasticity. It is checked by simple matching whether search time is reducible.

[0072]Search time measures time until search results are displayed after pushing the button for starting the search on an indicator 10 times manually, and makes the average value search time. CPU of having used it for the experiment this time is a personal computer Pentium4 (1.7 GHz) of U.S. Intel and whose main storage capacity are 654,812 KB.

[0073][F]The result of having compared the average of the average search length by the 2nd experiment as a result of the 2nd experiment is shown in drawing 15. x in a figure means that conformity results were not able to be searched. Although the result of the method by this invention is a little less in the songs C and G, by all the results, the result more than equivalent is obtained from drawing 15. In the method by this invention not using nonlinear matching, it turns out that it can be searching the conventional method and more than equivalent.

[0074]Next, using the song B in drawing 15 as a search key, search was repeated once and it compared by making the average into search time. In the conventional method, the place which had taken 4.29 seconds became about 2.42 seconds quick by the method by this invention, and shortening of the search time by simple matching was checked. That is, when how to use the conventional nonlinear elasticity was compared with the method of using linearity elasticity of this invention, while retrieval precision was equal in the case of the method by this invention, it turned out that processing speed (CPU burden) improves about 56%.

[0075]About the validity of the above this invention, it is clear that the same may be fundamentally said of Embodiments 1 and 4.

[0076]By the computer and a software program, can realize processing of each embodiment described above and the program, It can store in suitable recording media, such as a portable medium memory which a computer can read, semiconductor memory, and a hard disk, and a computer can be performed by reading from there. The program can be downloaded from other computers via a communication line, and can also install and perform it.

[0077]

[Effect of the Invention]Although dynamic matching which can respond to nonlinear elasticity of an audio signal was used for the distance calculation which expresses similarity with the method of searching the high audio signal section of similarity out of the audio signal of a retrieval object using the audio signal of the conventional search key, This invention reduces computational complexity compared with the case where an audio signal mainly uses dynamic matching for distance calculation by using static matching only corresponding to linearity elasticity for the distance calculation showing similarity linearity elasticity, however when there is nothing, and has the effect of lessening search time.

[0078]Even if all the audio signal sections that can use for a search key the audio signal section obtained as search results in this invention, and suit the retrieval object in the audio signal of a retrieval object by search of eyes once are not obtained, Search narrowed down by rechoosing the audio signal section which suits a retrieval object as a search key from the audio signal sections of search results can be performed, and it has the effect that a possibility that the audio signal which suits a retrieval object can be acquired becomes high.

* NOTICES *

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

[Drawing 1]It is a lineblock diagram for describing the embodiment of the invention 1.

[Drawing 2]It is a flow chart explaining operation of a voice storage phase and a voice time series feature amount vector extraction phase.

[Drawing 3]It is a figure explaining processing of a time window logging part.

[Drawing 4]It is a figure explaining the example of how to start a time window.

[Drawing 5]It is a figure explaining the linearity elastic method of partial tone voice time series characteristic quantity.

[Drawing 6]It is a flow chart explaining operation of a voice search phase and a search key voice time series feature amount vector extraction phase.

[Drawing 7]It is a lineblock diagram for describing the embodiment of the invention 2.

[Drawing 8]It is a flow chart explaining operation of a voice storage phase and a voice time series feature amount vector extraction phase.

[Drawing 9]It is a figure explaining how to carry out linearity compression of the partial tone voice time

series characteristic quantity of base period window length in a time base direction.

[Drawing 10]It is a flow chart explaining operation of a voice search phase and a search key voice time series feature amount vector extraction phase.

[Drawing 11]It is a lineblock diagram for describing the embodiment of the invention 3.

[Drawing 12]It is a lineblock diagram for describing the embodiment of the invention 3.

[Drawing 13]It is a flow chart explaining operation of a voice search phase.

[Drawing 14]It is a figure showing the result of the 1st experiment.

[Drawing 15]It is a figure showing the result of the 2nd experiment.

[Description of Notations]

P1 Voice storage phase

P2 Voice time series feature amount vector extraction phase

P3 Voice search phase

P4 search-key voice time series feature amount vector extraction phase

10 Audio signal retrieval device

11 Voice time series feature quantity extracting part

12 Time window logging part

13 Partial tone voice time series characteristic quantity fixed length-ized part

14 Voice storage part

15 Memory storage

16 Characteristic quantity information comparing element

17 Display style generation part

18 Search key voice time series feature quantity extracting part

20 Retrieval object sound signal input device

21 Display

22 Search key sound signal input device

23 Search condition input device

30 Voice time series characteristic quantity linearity compression zone

31 Search key voice time series characteristic quantity linearity compression zone

40 Search condition input section

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2003－248494
（P2003－248494A）
(43)公開日　平成15年９月５日(2003.9.5)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　11/00		Ｇ０６Ｆ　17/30	１７０Ｅ	５Ｂ０７５
Ｇ０６Ｆ　17/30	１７０		３５０Ｃ	５Ｄ０１５
	３５０	Ｇ１０Ｌ　3/00	５１５Ａ	
Ｇ１０Ｌ　15/00			５５１Ｇ	
15/02			５３１Ａ	

審査請求　未請求　請求項の数13　ＯＬ　（全 13 頁）　最終頁に続く

(21)出願番号　特願2002－47806(P2002－47806)

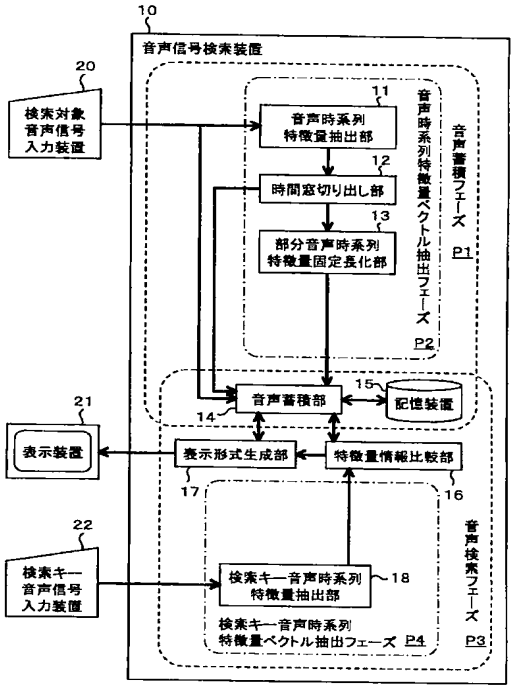
(22)出願日　平成14年２月25日(2002.2.25)

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　須賀　啓敏
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　寺本　純司
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(74)代理人　100087848
弁理士　小笠原　吉義　（外２名）

最終頁に続く

(54)【発明の名称】　音声信号検索方法，音声検索のための音声信号蓄積方法，音声信号検索装置，そのプログラムおよびそのプログラムの記録媒体

(57)【要約】
【課題】　音声データの検索において，固定長データ間の単純なマッチングにより類似距離の計算時間を短くし，検索キーと類似する音声データを高速に検索できるようにする。
【解決手段】　時間窓切り出し部１２により検索対象となる音声信号から複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出し，部分音声時系列特徴量固定長化部１３により部分音声時系列特徴量を線形伸縮して，所定の基準時間窓の長さにそろえ，これを音声蓄積部１４により蓄積する。検索時には，検索キーの入力音声信号から検索キー音声時系列特徴量抽出部１８により前記基準時間窓の長さの検索キー音声時系列特徴量ベクトルを抽出し，これと検索対象の音声時系列特徴量ベクトルとの類似距離計算を特徴量情報比較部１６で行って，類似度の高いものを検索結果とする。



1

【特許請求の範囲】

【請求項１】　検索対象となる音声信号から複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出する過程と，抽出された複数種類の長さの部分音声時系列特徴量を線形伸縮して，検索時における類似距離計算の基準となる基準時間窓の長さにそろえる過程と，前記基準時間窓の長さにそろえた部分音声時系列特徴量を，検索キーとして入力される音声信号から得られる音声時系列特徴量ベクトルと比較するための検索対象となる音声時系列特徴量ベクトルとして蓄積する過程と，検索キーとして入力または指定された音声信号から前記基準時間窓の長さの検索キー音声時系列特徴量ベクトルを抽出する過程と，前記検索対象として蓄積された音声時系列特徴量ベクトルと前記検索キー音声時系列特徴量ベクトルとの類似距離を計算し，検索キーの音声信号と検索対象である各音声信号区間の音声信号との類似度を算出する過程と，類似度の算出結果に基づいて検索結果を出力する過程とを有することを特徴とする音声信号検索方法。

【請求項２】　検索対象となる音声信号から，検索時における類似距離計算の基準となる長さの基準時間窓を使って部分音声時系列特徴量を抽出する過程と，抽出された部分音声時系列特徴量を，検索キーとして入力される音声信号から得られる音声時系列特徴量ベクトルと比較するための検索対象となる音声時系列特徴量ベクトルとして蓄積する過程と，検索キーとして入力または指定された音声信号から複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出する過程と，抽出された複数種類の長さの部分音声時系列特徴量を線形伸縮して，前記基準時間窓の長さにそろえる過程と，前記基準時間窓の長さにそろえた部分音声時系列特徴量を検索キー音声時系列特徴量ベクトルとし，前記検索対象として蓄積された音声時系列特徴量ベクトルと前記検索キー音声時系列特徴量ベクトルとの類似距離を計算して，検索キーの音声信号と検索対象である各音声信号区間の音声信号との類似度を算出する過程と，類似度の算出結果に基づいて検索結果を出力する過程とを有することを特徴とする音声信号検索方法。

【請求項３】　前記音声信号から複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出する過程では，前記音声信号から音声時系列特徴量を抽出し，そこから前記基準となる長さの基準時間窓を少しずつずらしながら部分音声時系列特徴量を切り出し，さらに基準時間窓の長さを中心とした複数種類の長さの時間窓でも同様に少しずつずらしながら部分音声時系列特徴量を切り出すことを特徴とする請求項１または請求項２記載の音声信号検索方法。

【請求項４】　前記検索対象として蓄積された音声時系列特徴量ベクトルと前記検索キー音声時系列特徴量ベクトルとは，基準時間窓の長さの部分音声時系列特徴量を

2

所定の長さに線形圧縮したものであることを特徴とする請求項１，請求項２または請求項３記載の音声信号検索方法。

【請求項５】　前記検索キーとして，一旦検索を行った後に検索結果として出力された音声信号区間を指定する情報を入力し，指定された音声信号区間に対応する音声時系列特徴量ベクトルを前記検索キー音声時系列特徴量ベクトルとして，前記検索対象として蓄積された音声時系列特徴量ベクトルとの類似距離計算により再検索を行うことを特徴とする請求項１から請求項４までのいずれかに記載の音声信号検索方法。

【請求項６】　検索対象となる音声信号を入力する過程と，入力された音声信号から音声時系列特徴量を抽出する過程と，前記音声時系列特徴量から基準となる長さの時間窓である基準時間窓を少しずつずらしながら部分音声時系列特徴量を切り出す過程と，前記基準時間窓の長さを中心とした複数種類の長さの時間窓についても少しずつずらしながら部分音声時系列特徴量を切り出す過程と，前記複数の種類の長さの部分音声時系列特徴量を線形伸縮して基準時間窓の長さにそろえる過程と，前記基準時間窓の長さにそろえた部分音声時系列特徴量を，検索キーとして入力される音声信号から得られる音声時系列特徴量ベクトルと比較するための検索対象となる音声時系列特徴量ベクトルとして蓄積する過程とを有することを特徴とする音声検索のための音声信号蓄積方法。

【請求項７】　前記基準時間窓の長さにそろえた部分音声時系列特徴量を，所定の長さに線形圧縮したものを，蓄積する検索対象の音声時系列特徴量ベクトルとすることを特徴とする請求項６記載の音声検索のための音声信号蓄積方法。

【請求項８】　検索対象となる音声信号から複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出する手段と，抽出された複数種類の長さの部分音声時系列特徴量を線形伸縮して，検索時における類似距離計算の基準となる基準時間窓の長さにそろえる手段と，前記基準時間窓の長さにそろえた部分音声時系列特徴量を，検索キーとして入力される音声信号から得られる音声時系列特徴量ベクトルと比較するための検索対象となる音声時系列特徴量ベクトルとして蓄積する手段と，検索キーとして入力または指定された音声信号から前記基準時間窓の長さの検索キー音声時系列特徴量ベクトルを抽出する手段と，前記検索対象として蓄積された音声時系列特徴量ベクトルと前記検索キー音声時系列特徴量ベクトルとの類似距離を計算し，検索キーの音声信号と検索対象である各音声信号区間の音声信号との類似度を算出する手段と，類似度の算出結果に基づいて検索結果を出力する手段とを備えることを特徴とする音声信号検索装置。

【請求項９】　検索対象となる音声信号から，検索時における類似距離計算の基準となる長さの基準時間窓を使

って部分音声時系列特徴量を抽出する手段と，抽出された部分音声時系列特徴量を，検索キーとして入力される音声信号から得られる音声時系列特徴量ベクトルと比較するための検索対象となる音声時系列特徴量ベクトルとして蓄積する手段と，検索キーとして入力または指定された音声信号から複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出する手段と，抽出された複数種類の長さの部分音声時系列特徴量を線形伸縮して，検索時における類似距離計算の基準となる基準時間窓の長さにそろえる手段と，前記基準時間窓の長さにそろえた部分音声時系列特徴量を検索キー音声時系列特徴量ベクトルとし，前記検索対象として蓄積された音声時系列特徴量ベクトルと前記検索キー音声時系列特徴量ベクトルとの類似距離を計算して，検索キーの音声信号と検索対象である各音声信号区間の音声信号との類似度を算出する手段と，類似度の算出結果に基づいて検索結果を出力する手段とを備えることを特徴とする音声信号検索装置。

【請求項１０】　請求項１から請求項５までのいずれかに記載の音声信号検索方法を，コンピュータに実行させるための音声信号検索用プログラム。

【請求項１１】　請求項１から請求項５までのいずれかに記載の音声信号検索方法を，コンピュータに実行させるためのプログラムを記録したことを特徴とする音声信号検索用プログラムの記録媒体。

【請求項１２】　請求項６または請求項７記載の音声検索のための音声信号蓄積方法を，コンピュータに実行させるための音声信号蓄積用プログラム。

【請求項１３】　請求項６または請求項７記載の音声検索のための音声信号蓄積方法を，コンピュータに実行させるためのプログラムを記録したことを特徴とする音声信号蓄積用プログラムの記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は，音声検索システムの技術に関し，特に入力または指定された音声信号に類似した音声信号区間を音声データベースの中から検索するための音声信号検索方法，音声検索のための音声信号蓄積方法，音声信号検索装置，そのプログラムおよびそのプログラムの記録媒体に関するものである。

【０００２】

【従来の技術】入力音声信号に類似した音声信号を音声データベースの中から検索する手法の従来技術としては，次の参考文献１に述べられたものがある。

［参考文献１］遠藤隆，中沢正幸，高橋裕信，岡隆一：“音声と動画像の自己組織化ネットワークによるデータ表現とスポッティング相互検索”：１９９８年度人工知能学会全国大会（第１２回），S5-04，pp.122-125.

これは，ＩＰＭ(Incremental Path Method) ネットワークを用いて，ＤＰマッチングと同様の動的マッチングで音声信号間の類似度を計算し，入力音声信号に類似した音声信号を音声データベースの中から検索する方法である。この方法によれば，動的マッチングを用いることにより，時間軸方向に非線形伸縮している音声信号でも類似度を計算することができる。

【０００３】

【発明が解決しようとする課題】音声信号を音声信号で検索しようとする場合，音声信号間の類似度を計算する必要がある。一般的に音声信号は，同一文字列を発声した場合でも時間軸上で非線形伸縮する。このことから，音声信号間の類似度を計算するには，非線形伸縮に対応するため，ＤＰマッチングや隠れマルコフモデル（HMM）等の動的マッチングを用いる必要があった。

【０００４】しかし，全ての音声信号が時間軸上で非線形伸縮するわけではない。非線形伸縮しない音声信号の例として，まず歌声が挙げられる。歌声には潜在的にテンポが存在するので，短時間区間であれば，その区間全体のテンポがゆっくりになったり，速くなったりはすることはあるものの，その区間内でテンポが乱れることはない。つまり歌声は短時間区間内であれば，その区間全体で線形伸縮することはあるものの，その区間内で非線形伸縮することはない。

【０００５】また，非線形伸縮しない音声信号の他の例として，アナウンサーの発話音声がある。アナウンサーは同じ言葉を同じ調子で何度も発話することが非常に上手であるため，同一の言葉であれば非線形伸縮しない。さらに，一般の人の発話でも一単語程度の短い発話であれば，その区間全体で線形伸縮することはあっても，その区間内では，ほとんど非線形伸縮していないとみなすことができる。

【０００６】従来の技術では，音声信号間の類似度計算をするときに動的マッチングを用いていた。したがって，線形伸縮だけしかしていない音声にも，非線形伸縮にも対応できる動的マッチングを行ってしまう。動的マッチングは，固定長ベクトル間のユークリッド距離やマンハッタン距離等の距離計算を行う静的マッチングに比べて計算量が多く，検索時間が長くなってしまうという問題点があった。

【０００７】

【課題を解決するための手段】前記課題を解決するために，本発明は，音声データの検索において時間窓を用いて切り出した特徴量を動的マッチングではなく，静的マッチングによって高速に照合できるようにする。そのため，音声時系列特徴量の音声データを線形伸縮させて，ある一定の時間区間の長さの固定長データにそろえる。こうすることにより，固定長データ間の単純なマッチングによって，検索キーと類似する音声データを高速に検索可能にすることを特徴とする。

【０００８】具体的には，本発明の音声信号検索装置は，音声時系列特徴量抽出手段と，時間窓切り出し手段

と，部分音声時系列特徴量固定長化手段と，検索キー音声時系列特徴量抽出手段と，音声蓄積手段と，検索条件入力手段と，特徴量情報比較手段と，表示形式生成手段からなる。さらに，音声時系列特徴量線形圧縮手段と，検索キー音声時系列特徴量線形圧縮手段とを設けてもよい。

【０００９】音声時系列特徴量抽出手段は，時系列信号である音声信号から音声時系列特徴量を抽出する。

【００１０】時間窓切り出し手段は，まず基準となる長さの時間窓である基準時間窓を用意する。そして，音声時系列特徴量抽出手段により抽出された音声時系列特徴量から，基準時間窓の長さの部分音声時系列特徴量を基準時間窓を少しずつずらしながら切り出す。次に，基準時間窓の長さを中心にして複数種類の長さの時間窓を用意する。これらの長さの時間窓でも同様にして部分音声時系列特徴量を切り出す。

【００１１】部分音声時系列特徴量固定長化手段は，時間窓切り出し手段により複数種類の長さの時間窓で切り出した部分音声時系列特徴量を，線形伸縮して基準時間窓長にそろえた部分音声時系列特徴量を生成して，それを音声時系列特徴量ベクトルとして抽出する。

【００１２】音声時系列特徴量線形圧縮手段は，部分音声時系列特徴量固定長化手段により生成された基準時間窓長の音声時系列特徴量を，ある一定の長さに線形圧縮して音声時系列特徴量ベクトルを抽出する。

【００１３】検索キー音声時系列特徴量抽出手段は，検索キーとして入力された基準時間窓の長さの音声信号から，基準時間窓長の検索キー音声系列特徴量を抽出して，それを検索キー音声時系列特徴量ベクトルとして抽出する。

【００１４】検索キー音声時系列特徴量線形圧縮手段は，検索キー音声時系列特徴量抽出手段により抽出された基準時間窓長の音声時系列特徴量を，音声時系列特徴量ベクトルと同じ長さに線形圧縮して，検索キー音声時系列特徴量ベクトルを抽出する。

【００１５】音声蓄積手段は，入力された検索対象の音声信号を蓄積し，また音声時系列特徴量線形圧縮手段により抽出された音声時系列特徴量ベクトルからインデックスを作成して音声時系列特徴量ベクトルとともに蓄積し，さらに音声時系列特徴量ベクトルの抽出元の音声信号区間への対応付けの情報である時間窓切り出し情報も蓄積する。

【００１６】検索条件入力手段は，音声蓄積手段において蓄積された音声信号区間を検索キーとする際に，検索キーを指定するための条件を入力する。

【００１７】特徴量情報比較手段は，検索キー音声時系列特徴量線形圧縮手段において抽出された検索キー音声時系列特徴量ベクトルと音声蓄積手段において蓄積された音声時系列特徴量ベクトルを静的マッチングで類似距離を計算し，類似度を設定する。これにより動的マッチングよりも高速な類似度計算が可能になる。そして，検索キー音声時系列特徴量ベクトルとの類似度が高い順に音声時系列音声特徴量ベクトルを出力する。

【００１８】表示形式生成手段は，特徴量情報比較手段から順位付けられて出力された音声時系列特徴量ベクトルを音声蓄積手段に蓄積された時間窓切り出し情報を元に抽出元の音声信号区間に対応付けて表示装置に出力する。

【００１９】

【発明の実施の形態】〔実施の形態１〕図１は，本発明の実施の形態１を説明するための構成図である。実施の形態１では，外部から入力した音声を検索キーとして用いる。

【００２０】本発明の動作は，音声蓄積フェーズＰ１と，それから呼び出される音声時系列特徴量ベクトル抽出フェーズＰ２と，音声検索フェーズＰ３と，それから呼び出される検索キー音声時系列特徴量ベクトル抽出フェーズＰ４とで構成される。以下，各フェーズの動作を説明する。

【００２１】〔Ａ〕音声蓄積フェーズＰ１と音声時系列特徴量ベクトル抽出フェーズＰ２

図２は，音声蓄積フェーズＰ１と音声時系列特徴量ベクトル抽出フェーズＰ２の動作を説明するフローチャートである。

【００２２】まず，検索対象音声信号入力装置２０から検索対象の音声信号が入力され，音声蓄積部１４においてこの音声信号を記憶装置１５に蓄積する（ステップＳ１）。

【００２３】次に，音声時系列特徴量抽出部１１において，入力された音声信号から音声時系列特徴量を抽出する（ステップＳ２）。音声時系列特徴量としては，例えば，メル周波数ケプストラム係数の低次項や，その１次差分，２次差分や，音声パワーや，フィルタバンク分析による各帯域の音声パワー等を多次元ベクトルで表し，それらを時系列順に並べたものを用いることができる。音声時系列特徴量の例は，次の参考文献２に述べられている。

［参考文献２］鹿野清宏他：“ＩＴ　ｔｅｘｔ　音声認識システム”，オーム社，2001.

次に，時間窓切り出し部１２において，まず図３に示すように，基準となる長さの時間窓を基準時間窓として設定し，この基準時間窓を少しずつずらしながら，基準時間窓長の部分音声時系列特徴量を切り出す（ステップＳ３）。

【００２４】図４に示すように，基準時間窓の長さを中心とした複数種類の長さの時間窓を設定し，基準時間窓の場合と同様に，時間窓を少しずつずらしながらその時間窓長の部分音声時系列特徴量をそれぞれ切り出す（ステップＳ４）。後述する実験例では，基準時間窓は１５０フレームであり，１フレーム当たり約２６ミリ秒の長

さである。時間窓の長さの下限は１１８フレーム，上限は１８２フレームとした。これらの時間窓を使って部分音声時系列特徴量を切り出したならば，その時間窓の切り出しに関する情報を音声蓄積部１４において時間窓切り出し情報として記憶装置１５に蓄積しておく。

【００２５】また，部分音声時系列特徴量固定長化部１３において，図５に示すように，複数種類の長さの時間窓で切り出された部分音声時系列特徴量をそれぞれ時間軸方向で線形伸縮させて基準時間窓の長さにそろえ（ステップＳ５），その基準時間窓長の部分音声時系列特徴量を生成して，それを音声時系列特徴量ベクトルとする（ステップＳ６）。

【００２６】そして，得られた音声時系列特徴量ベクトルからインデックスを構築し，音声時系列特徴量ベクトルとともに音声蓄積部１４の記憶装置１５に蓄積する（ステップＳ７）。多次元空間ベクトルのインデックス構造としては，例えば，次の参考文献３に述べられているＳＲ－ｔｒｅｅや，参考文献４に述べられているＡ－ｔｒｅｅなどを利用することができる。

［参考文献３］Norio Katayama and Shin'ichi Satoh：“The SR-Tree ：An Index Structure for High-Dimensional Nearest Neighbor Queries”，In Proc. ACM SIGMOID International Conference on Management of Data，pp.368-380，May 1997.

［参考文献４］Yasushi Sakurai，Masatoshi Yoshikawa，Shunsuke Uemura，and Haruhiko Kojima：“The A-Tree：An Index Structure for High-Dimensional Spaces Using Relative Approximation”，In Proc. of the 26th International Conference on Very Large Data Bases（VLDB），pp.516-526，Cairo，September 2000.

〔Ｂ〕音声検索フェーズＰ３と検索キー音声時系列特徴量ベクトル抽出フェーズＰ４

図６は，音声検索フェーズＰ３と検索キー音声時系列特徴量ベクトル抽出フェーズＰ４の動作を説明するフローチャートである。

【００２７】まず，検索キー音声信号入力装置２２を用いて，検索キーとなる基準時間窓長の音声信号を入力する（ステップＳ１０）。

【００２８】次に，検索キー音声時系列特徴量抽出部１８において，入力した基準時間窓長の音声信号から，基準時間窓長の検索キー音声時系列特徴量を抽出する（ステップＳ１１）。検索キー音声時系列特徴量としては，前述した音声蓄積フェーズＰ１の音声時系列特徴量と同じ特徴量を利用する。この検索キー音声時系列特徴量を検索キー音声時系列特徴量ベクトルとする（ステップＳ１２）。

【００２９】さらに，特徴量情報比較部１６において，得られた検索キー音声時系列特徴量ベクトルと，音声蓄積部１４において記憶装置１５に蓄積された音声時系列特徴量ベクトルとの類似距離を計算する（ステップＳ１３）。この距離計算には，固定長ベクトル間のユークリッド距離やマンハッタン距離等の静的マッチングを用い，また音声蓄積部１４の記憶装置１５に蓄積されたインデックスを用いて行う。これにより動的マッチングよりも高速な距離計算が可能となる。そして，その距離の短い順に音声時系列特徴量ベクトルの音声信号区間を順位付ける（ステップＳ１４）。

【００３０】最後に，表示形式生成部１７において，音声蓄積部１４により記憶装置１５に蓄積された時間窓切り出し情報を用いて音声時系列特徴量ベクトルをその抽出元の音声信号区間に対応付けて，表示装置２１に出力する（ステップＳ１５）。ここでの表示では，例えば１時間の音楽番組の中から検索キーに該当する部分の検索結果を表示する場合に，順位の高い結果から順番に，番組の先頭から何分何秒目であるかなどの音声信号区間を示す一覧情報と，その部分の再生用のボタンとを表示し，再生ボタンが押された場合にはその部分の音声を出力することを行う。これによって，検索者が検索目的に適合する音声信号区間を探し出す手間を省けるようになる。また，検索対象について曲名などの情報をデータベース中に持つ場合には，検索結果の曲名などを併せて表示することができる。

【００３１】〔実施の形態２〕図７は，本発明の実施の形態２を説明するための構成図である。実施の形態２は，時間窓を用いて切り出した特徴量を圧縮してサイズを小さくし，サイズのより小さいデータベースを構築して検索を行うことを可能にしたものである。

【００３２】実施の形態２では，線形伸縮させて基準時間窓の長さにそろえた部分音声時系列特徴量を時間軸上で線形圧縮したものを音声時系列特徴量ベクトルとして用い，これに伴い，基準時間窓の長さの検索キー音声時系列特徴量を音声時系列特徴量ベクトルと同様の長さに時間軸上で線形圧縮したものを検索キー音声時系列特徴量ベクトルとして用いる。そのため実施の形態１と比べて，音声蓄積フェーズＰ１から呼び出される音声時系列特徴量ベクトル抽出フェーズＰ２と，音声検索フェーズＰ３から呼び出される検索キー音声時系列特徴量ベクトル抽出フェーズＰ４が異なり，以下の通りになる。

【００３３】〔Ａ〕音声蓄積フェーズＰ１と音声時系列特徴量ベクトル抽出フェーズＰ２

図８は，音声蓄積フェーズＰ１と音声時系列特徴量ベクトル抽出フェーズＰ２の動作を説明するフローチャートである。

【００３４】実施の形態１と同様に，検索対象音声信号入力装置２０から音声信号を入力する（ステップＳ２０）。音声時系列特徴量抽出部１１において，入力された音声信号から音声時系列特徴量を抽出し（ステップＳ２１），時間窓切り出し部１２において，基準時間窓を少しずつずらしながら，基準時間窓長の部分音声時系列

特徴量を切り出す（ステップＳ２２）。さらに，複数種類の長さの時間窓でも部分音声時系列特徴量を切り出し（ステップＳ２３），部分音声時系列特徴固定長化部１３において，切り出された複数種類の長さの部分音声時系列特徴量を時間軸上で線形伸縮して，基準時間窓長にそろえた部分音声時系列特徴量を生成する（ステップＳ２４）。

【００３５】次に，音声時系列特徴量線形圧縮部３０において，図９に示すように，基準時間窓の長さにそろえた部分音声時系列特徴量をそれぞれ時間軸方向に線形圧縮して，音声時系列特徴量ベクトルを抽出する（ステップＳ２５）。これにより，音声時系列特徴量ベクトルの次元数が小さくなり，類似度計算の計算量を減らすことができ，蓄積する記憶装置１５の容量を小さくすることもできる。

【００３６】さらに，実施の形態１と同様にして，得られた音声時系列特徴量ベクトルからインデックスを構築し，音声時系列特徴量ベクトルとともに音声蓄積部１４の記憶装置１５に蓄積する。

【００３７】〔Ｂ〕音声検索フェーズＰ３と検索キー音声時系列特徴量ベクトル抽出フェーズＰ４

図１０は，音声検索フェーズＰ３と検索キー音声時系列特徴量ベクトル抽出フェーズＰ４の動作を説明するフローチャートである。

【００３８】実施の形態１と同様に，検索キー音声信号入力装置２２を用いて，検索キーとなる基準時間窓長の音声信号を入力し（ステップＳ３０），検索キー音声時系列特徴量抽出部１８において，入力した基準時間窓長の音声信号から，基準時間窓長の検索キー音声時系列特徴量を抽出する（ステップＳ３１）。

【００３９】そして，検索キー音声時系列特徴量線形圧縮部３１において，図９に示すように，基準時間窓長の音声時系列特徴量を時間軸方向に線形圧縮し，検索キー音声時系列特徴量ベクトルを抽出する（ステップＳ３２）。なお，検索キー音声時系列特徴量ベクトルの長さは，前記音声蓄積フェーズＰ１で生成された音声時系列特徴量ベクトルと同じ長さとする。

【００４０】さらに，実施の形態１と同様に，特徴量情報比較部１６において，得られた検索キー音声時系列特徴量ベクトルと，音声蓄積部１４において記憶装置１５に蓄積された音声時系列特徴量ベクトルとの類似距離を計算し（ステップＳ３３），その距離の短い順に音声時系列特徴量ベクトルを順位付ける（ステップＳ３４）。

【００４１】最後に，実施の形態１と同様に，表示形式生成部１７において，音声蓄積部１４の記憶装置１５に蓄積された時間窓切り出し情報を用いて音声時系列特徴量ベクトルをその抽出元の音声信号区間に対応付けて，表示装置２１に出力する（ステップＳ３５）。例えば１時間の音楽番組の中から検索キーに該当する部分の検索結果を表示する場合に，順位の高い結果から順番に，番組の先頭から何分何秒目であるかなどの音声信号区間を示す一覧情報と，その部分の再生用のボタンとを表示し，再生ボタンが押された場合にはその部分の音声を出力することができるような表示形式で表示する。

【００４２】〔実施の形態３〕図１１，図１２は，本発明の実施の形態３を説明するための構成図である。実施の形態３は，検索キーを外部から入力するのではなく，一度検索を行って検索結果を得たときに検索結果の中から新たに検索キーを指定しなおして，それに類似する音声データを検索することができるようにしたものである。

【００４３】音声データを検索キーとする複数の検索結果を類似度の順に一覧表のように画面に表示し，検索結果として得られた音声データを新たな検索キーとして，類似する他の音声データを検索する。

【００４４】例えば音楽検索の場合，検索結果表示画面の「曲名」の部分に触れると，その楽曲が検索結果として選択され，検索キーに対応する部分または楽曲の先頭から再生されて音声出力される。また，画面上の「順位」の部分に触れると，その楽曲の検索結果（例えば楽曲のうち当初の検索キーである音声データと類似する部分（３～４秒間））を新たな検索キーとして，さらに類似する楽曲の検索を実行するようなことを行う。

【００４５】以上のように実施の形態３では，検索結果の中から，検索目的に最も近かった結果を検索キーとして選択して再び検索を行う。図１１に示す音声信号検索装置１０の音声蓄積フェーズＰ１は，図１に示す実施の形態１の音声蓄積フェーズＰ１と同様であり，図１２に示す音声信号検索装置１０の音声蓄積フェーズＰ１は，図７に示す実施の形態２の音声蓄積フェーズＰ１と同様である。前述した実施の形態１，実施の形態２とは，音声検索フェーズＰ３が異なり，以下のようになる。

【００４６】〔Ａ〕音声検索フェーズＰ３

図１３は，音声検索フェーズＰ３の動作を説明するフローチャートである。

【００４７】前段階として，実施の形態１，実施の形態２と同様に検索を行い，その検索結果を類似度順に順位付けて表示装置２１に表示する（ステップＳ４０）。ユーザからの指示により，表示された検索結果が検索目的に十分に合致していれば検索を終了する（ステップＳ４１）。

【００４８】検索目的に合致していなければ，検索条件入力部２３において，既に表示されている順位付けられた検索結果のうちの検索目的に最も近い結果を検索キーとして指定させ，それを検索キーとして選びなおす（ステップＳ４２）。このため，ステップＳ４０において順位付けられた結果を表示装置２１で表示する際に，結果１件あたりに２つの入力用ボタンを表示する。１つは検索キーとして指定する際に押すボタンであり，１つはその結果の音声を発声させる際に押すボタンである。ユー

ザは，表示装置２１上で前者のボタンを押すことで，検索結果の中から検索キーを指定することができる。

【００４９】検索キーが指定されると，特徴量情報比較部１６において，検索キーと類似度の高い蓄積された音声区間を検索し，類似度の高い順に順位付けする（ステップＳ４３）。さらに，表示形式生成部１７において，検索結果から検索キーを選択できる表示形式を生成し，表示装置２１に表示する（ステップＳ４０）。

【００５０】表示された検索結果が，検索目的に十分に合致していれば検索を終了する（ステップＳ４１）。まだ十分に合致しない場合には，さらに検索キーを選択しなおして再び検索を行い，検索目的に十分に合致する結果が得られるまで繰り返す（ステップＳ４０～Ｓ４３）。

【００５１】〔実施の形態４〕以上の実施の形態１～３では，検索対象の音声信号について複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出し，それらの部分音声時系列特徴量を線形伸縮して基準時間窓の長さにそろえたものを，検索対象の音声時系列特徴量ベクトルとして，音声蓄積部１４により蓄積した。

【００５２】実施の形態４では，時間窓で切り出した部分音声時系列特徴量の線形伸縮を，検索対象のものについて行うのではなく，検索キーとして入力されたものについて行う。すなわち，実施の形態４では，検索対象の音声信号から複数種類の長さの時間窓を使って部分音声時系列特徴量を抽出するのではなく，基準時間窓だけを使って部分音声時系列特徴量を抽出し，それを検索対象の音声時系列特徴量ベクトルとして音声蓄積部１４に蓄積する。一方，検索キーとして入力された音声信号については，複数種類の長さの時間窓を使ってそれぞれ長さの異なる部分音声時系列特徴量を抽出し，これらの長さの異なる部分音声時系列特徴量を，基準時間窓の長さになるように線形伸縮する。

【００５３】線形伸縮を検索キーとして入力されたものについて行っても，実施の形態１～３と同様な検索結果が得られる。なお，基準時間窓の長さにそろえたのちに，必要に応じて実施の形態２のように一定の長さに線形圧縮することにより，検索時に照合するデータ量の削減を図ることもできる。

【００５４】〔実験結果〕本発明の有効性を確認するため，実施の形態２について実験用に作成した歌声データを対象とした２種類の実験を行った。

【００５５】まず第１の実験で，本発明による線形伸縮に対応したマッチングが，線形伸縮に対応しないマッチングに比べて，歌声の検索において有効であることを確認し，次に第２の実験で，非線形伸縮マッチングを実装している従来方式と比較しても，本発明が歌声の検索において十分に有効であることを確認した。

【００５６】〔Ａ〕両実験に共通する実験条件

１５人の被験者を女性，男性，混合の３つのグループに分け，それぞれのグループに６２曲の歌名リストを渡す。歌名リストの中で，その歌を知っている人にフレーズの一部（約１０秒程度）を歌ってもらい，６２×３個の歌声（合計約３０分）をデータベースに格納する。

【００５７】１つの被験者グループの歌声の中から任意に１２曲選び，そこから１フレーズ程度（基準時間窓長：１５０フレーム分，約４秒）を取り出して検索キーとする。そして，他の２つのグループの被験者の歌声の同一フレーズ部分を適合結果として検索する。

【００５８】本実験では，サンプリング周波数４４１００Ｈｚ，量子化ビット数１６ｂｉｔ，１チャネルのｗａｖｅファイル形式の音声データから歌声を検索することとする。音声特徴量として，メル周波数ケプストラム係数の低次項５次元を使い，これを時間軸上に並べたものから時間窓を用いて時系列特徴量を抽出する。

【００５９】検索結果の評価基準としては，平均探索長の平均を用いる。これは検索結果の中から，検索目的に適合する結果を探し出す手間を表す評価基準である。なお，順位付けされた検索結果のうち，２０位までの適合性を判断することとし，順位を２０位までとし，それ以下に適合結果があっても検索できなかったものとする。

【００６０】〔Ｂ〕平均探索長の平均の説明

ここで，検索結果の評価基準として用いた平均探索長の平均について説明する。平均探索長については，次の参考文献５に述べられている。

［参考文献５］徳永健伸：“言語と計算５　情報検索と言語処理”，東京大学出版，1999.

平均探索長は，検索結果として順位付けられた集合を評価する尺度である。検索結果として順位付けられた結果が返ってきた場合，実際には検索者は，検索結果の適合性を上位の結果から逐一判断していかなければいけない。平均探索長は，このような検索者の適合性判断の過程を考慮し，検索者が必要な数の適合結果を得るためには，どれだけ結果の適合性を判断しなければならないかというユーザの手間を計測する尺度である。

【００６１】例えば，検索結果が次のように順序付き集合S_1，S_2，S_3に分けることができたとする。ただし，集合間の順序はS_1，S_2，S_3の順であり，○，×はそれぞれ適合結果，不適合結果を表す。

【００６２】S_1：{○，×，×，×}

S_2：{○，○，○，○，×，×}

S_3：{○，○，×，×}

今，検索者が，適合結果を１つ得たいとする。まず，集合S_1を検査することになる。この集合の中では順序が付いていないので，適合結果を見つけるまでに検査しなければならない結果の個数の期待値は，

１×１／４＋２×１／４＋３×１／４＋４×１／４＝２．５

となる。

【００６３】これは検索結果から適合結果を１つ見つけるためには，検索者は平均的に２．５個の検索結果の適合性を判断しなければならないことを示している。つまり，この検索結果から１つの適合結果を見つけ出すのに必要な平均探索長は，２．５個である。

【００６４】また，適合結果を２つ見つけるためには，集合S_1を全部検査した後，集合S_2から１つ見つければよいから，検査すべき結果の個数の期待値は，

（４＋１）×４／６＋（４＋２）×４／１５＋（４＋３）×１／１５＝５．４

となる。つまり，２つの適合結果を見つけ出すのに必要な平均探索長は，５．４個である。

【００６５】上記の例では，検索結果が順位付けられた集合で与えられている場合であったが，検索結果の個々に全順序が付けられている場合でも，各集合の要素を１つと考えれば平均探索長を計算できる。

【００６６】以上のことからもわかるように，平均探索長は一つの尺度とはならず，必要な適合結果の個数に依存した値となる。そこで，平均探索長の平均として，必要な適合結果１つあたりの平均探索長の値を計算する。

【００６７】必要な適合結果数をｉ，総適合結果数をＭ，ｉ個の必要な適合結果を見つけるのに必要な平均探索長をｘ（ｉ）とすると，平均探索長の平均x_{av}は，

$x_{av}=(1/M)\Sigma_{i=1}^{M}\{x(i)/i\}$

で表される。

【００６８】例えば，検索の結果，適合した結果が２位と６位に検索された場合を考える。必要な適合結果の個数を１とした場合，平均探索長は２となり，必要な検索結果の個数を２とした場合，平均探索長は６となる。これらの平均探索長の平均は，（２／１＋６／２）／２＝４となる。

【００６９】〔Ｃ〕第１の実験

線形伸縮に対応しない方法として，１種類の長さの時間窓だけを使って時系列特徴量を抽出する方法で実験する。一方，線形伸縮に対応する方法として，９種類の時間窓を使った本発明による方法で実験する。この両方法の平均探索長の平均の値を比較する。

【００７０】〔Ｄ〕第１の実験の結果

第１の実験による平均探索長の平均を比較した結果を図１４に示す。図中の×は，適合結果を検索できなかったことを表している。平均探索長の平均は，歌Ｂ，Ｅ，Ｈで同じ値になるものの，それ以外の歌ではすべて本発明による方法が上回っている。すなわち，固定長の時間窓により検索対象の音声データを対象として，従来方法による線形伸縮を用いない場合と，本発明による線形伸縮を用いた場合とを比較すると，平均探索長が短くて済み，検索精度が２５％から２倍程度向上するという効果があることがわかった。したがって，本発明による線形伸縮に対応する方式は，歌声の検索において有効であると言える。

【００７１】〔Ｅ〕第２の実験

非線形伸縮に対応できる従来方法として，メディアドライブ株式会社の「CrossMediator for Video V2.0(R1)」のボイス検索機能を用いて，従来方法と本発明による方法の平均探索長の平均を比較する。また，単純なマッチングにより，検索時間を削減できているかも確認する。

【００７２】検索時間は，表示部上の検索を開始するためのボタンを押した後から検索結果が表示されるまでの時間を手動で１０回計測し，その平均値を検索時間とする。なお，今回実験に使用したのは，ＣＰＵが米国Ｉｎｔｅｌ社のＰｅｎｔｉｕｍ４（１．７ＧＨｚ），主記憶容量が６５４，８１２ＫＢのパーソナルコンピュータである。

【００７３】〔Ｆ〕第２の実験の結果

第２の実験による平均探索長の平均を比較した結果を図１５に示す。図中の×は，適合結果を検索できなかったことを表している。図１５から，歌Ｃ，Ｇでは，本発明による方法の結果が若干下回っているものの，その他すべての結果では同等以上の結果が得られている。非線形マッチングを使わない本発明による方法では，従来方法と同等以上に検索できていることがわかる。

【００７４】次に，検索キーとして図１５中の歌Ｂを用い，１回検索を繰り返し，その平均を検索時間として比較を行った。従来方法では，４．２９秒かかっていたところが，本発明による方法では２．４２秒ほど速くなり，単純なマッチングによる検索時間の短縮が確認された。すなわち，従来の非線形伸縮を用いる方法と，本発明の線形伸縮を用いる方法とを比較すると，本発明による方法の場合，検索精度は遜色がない一方，処理速度（ＣＰＵ負担）は５６％程度向上することがわかった。

【００７５】以上の本発明の有効性については，実施の形態１，４についても基本的に同様であることは明らかである。

【００７６】以上説明した各実施の形態の処理は，コンピュータとソフトウェアプログラムとによって実現することができ，そのプログラムは，コンピュータが読み取り可能な可搬媒体メモリ，半導体メモリ，ハードディスク等の適当な記録媒体に格納して，そこから読み出すことによりコンピュータに実行させることができる。また，そのプログラムは通信回線を経由して他のコンピュータからダウンロードすることができ，それをインストールして実行させることもできる。

【００７７】

【発明の効果】従来の検索キーの音声信号を用いて検索対象の音声信号の中から類似度の高い音声信号区間を検索する方法では，類似度を表す距離計算に音声信号の非線形伸縮に対応できる動的マッチングを用いていたが，本発明は，音声信号が主に線形伸縮しかしないような場合に，類似度を表す距離計算に線形伸縮だけに対応する静的マッチングを用いることで動的マッチングを距離計

算に用いる場合に比べて計算量を削減し，検索時間を少なくするという効果を有する。

【００７８】また，本発明では，検索キーに検索結果として得られた音声信号区間を利用することができ，一度目の検索で検索対象の音声信号中の検索目的に適合する全ての音声信号区間が得られなくても，検索結果の音声信号区間の中から検索目的に適合する音声信号区間を検索キーとして選びなおすことで絞り込んだ検索を行うことができ，検索目的に適合する音声信号を得られる可能性が高くなるという効果を有する。

【図面の簡単な説明】

【図１】本発明の実施の形態１を説明するための構成図である。

【図２】音声蓄積フェーズと音声時系列特徴量ベクトル抽出フェーズの動作を説明するフローチャートである。

【図３】時間窓切り出し部の処理を説明する図である。

【図４】時間窓の切り出し方法の例を説明する図である。

【図５】部分音声時系列特徴量の線形伸縮方法を説明する図である。

【図６】音声検索フェーズと検索キー音声時系列特徴量ベクトル抽出フェーズの動作を説明するフローチャートである。

【図７】本発明の実施の形態２を説明するための構成図である。

【図８】音声蓄積フェーズと音声時系列特徴量ベクトル抽出フェーズの動作を説明するフローチャートである。

【図９】基準時間窓長の部分音声時系列特徴量を時間軸方向に線形圧縮する方法を説明する図である。

【図１０】音声検索フェーズと検索キー音声時系列特徴量ベクトル抽出フェーズの動作を説明するフローチャートである。

【図１１】本発明の実施の形態３を説明するための構成図である。

【図１２】本発明の実施の形態３を説明するための構成図である。

【図１３】音声検索フェーズの動作を説明するフローチャートである。

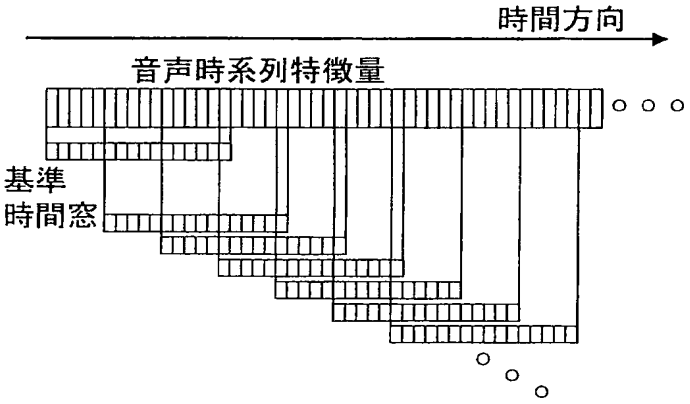
【図１４】第１の実験の結果を示す図である。

【図１５】第２の実験の結果を示す図である。

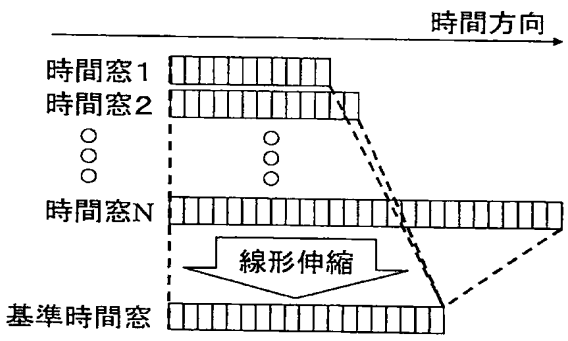
【符号の説明】

Ｐ１　音声蓄積フェーズ
Ｐ２　音声時系列特徴量ベクトル抽出フェーズ
Ｐ３　音声検索フェーズ
Ｐ４　検索キー音声時系列特徴量ベクトル抽出フェーズ
１０　音声信号検索装置
１１　音声時系列特徴量抽出部
１２　時間窓切り出し部
１３　部分音声時系列特徴量固定長化部
１４　音声蓄積部
１５　記憶装置
１６　特徴量情報比較部
１７　表示形式生成部
１８　検索キー音声時系列特徴量抽出部
２０　検索対象音声信号入力装置
２１　表示装置
２２　検索キー音声信号入力装置
２３　検索条件入力装置
３０　音声時系列特徴量線形圧縮部
３１　検索キー音声時系列特徴量線形圧縮部
４０　検索条件入力部

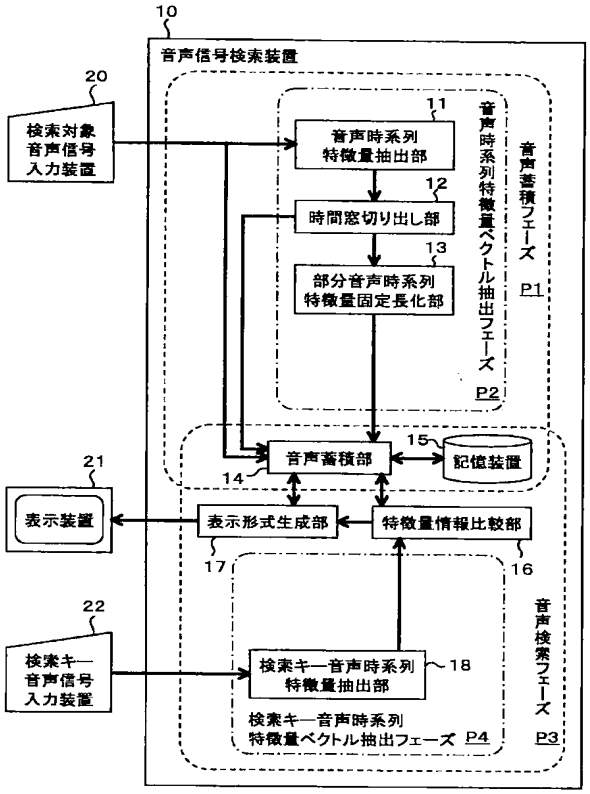
【図３】



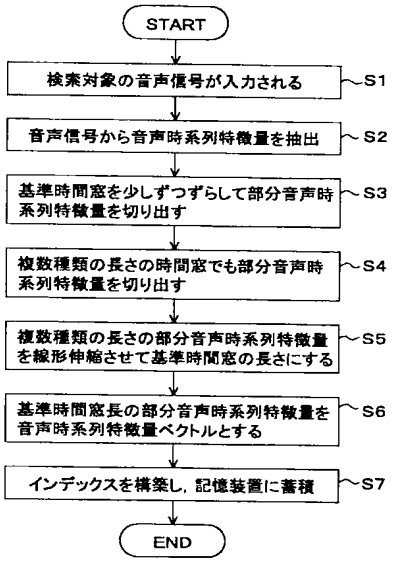
【図５】



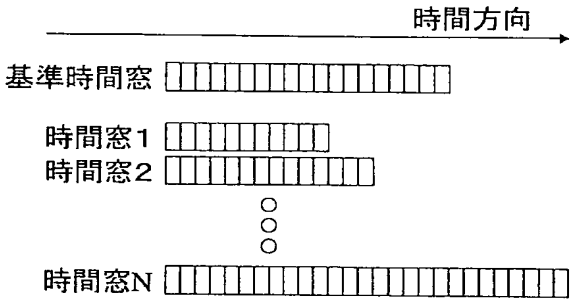
【図１】


10
音声信号検索装置
20
検索対象
音声信号
入力装置
11
音声時系列
特徴量抽出部
12
時間窓切り出し部
13
部分音声時系列
特徴量固定長化部
音声時系列特徴量ベクトル抽出フェーズ
P2
音声蓄積フェーズ
P1
15
記憶装置
音声蓄積部
14
21
表示装置
表示形式生成部
特徴量情報比較部
17
16
22
検索キー
音声信号
入力装置
検索キー音声時系列
特徴量抽出部
18
検索キー音声時系列
特徴量ベクトル抽出フェーズ P4
音声検索フェーズ
P3

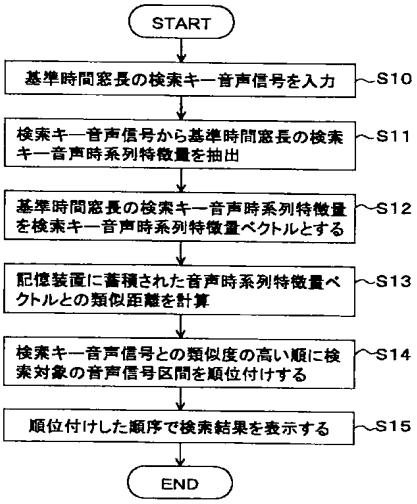
【図２】


START
検索対象の音声信号が入力される
S1
音声信号から音声時系列特徴量を抽出
S2
基準時間窓を少しずつずらして部分音声時系列特徴量を切り出す
S3
複数種類の長さの時間窓でも部分音声時系列特徴量を切り出す
S4
複数種類の長さの部分音声時系列特徴量を線形伸縮させて基準時間窓の長さにする
S5
基準時間窓長の部分音声時系列特徴量を音声時系列特徴量ベクトルとする
S6
インデックスを構築し，記憶装置に蓄積
S7
END

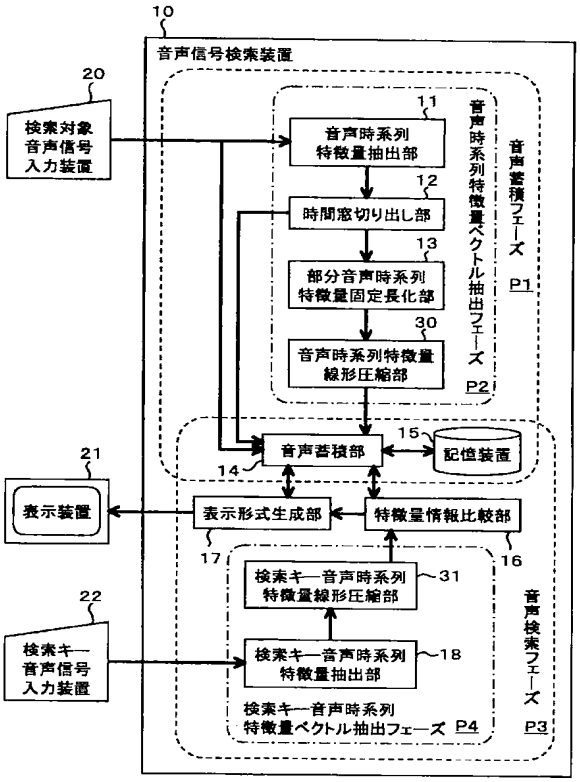
【図４】


時間方向
基準時間窓
時間窓１
時間窓２
時間窓N

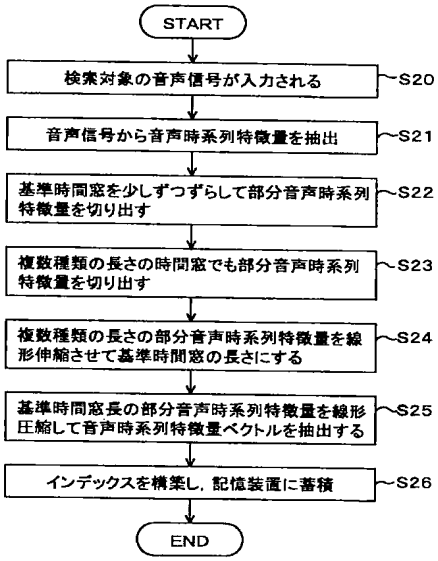
【図６】


START
基準時間窓長の検索キー音声信号を入力
S10
検索キー音声信号から基準時間窓長の検索キー音声時系列特徴量を抽出
S11
基準時間窓長の検索キー音声時系列特徴量を検索キー音声時系列特徴量ベクトルとする
S12
記憶装置に蓄積された音声時系列特徴量ベクトルとの類似距離を計算
S13
検索キー音声信号との類似度の高い順に検索対象の音声信号区間を順位付けする
S14
順位付けした順序で検索結果を表示する
S15
END

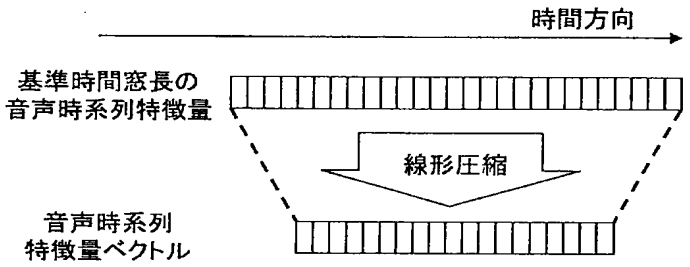
【図７】


10
音声信号検索装置
20
検索対象音声信号入力装置
11
音声時系列特徴量抽出部
12
時間窓切り出し部
13
部分音声時系列特徴量固定長化部
30
音声時系列特徴量線形圧縮部
音声時系列特徴量ベクトル抽出フェーズ P2
音声蓄積フェーズ P1
15
記憶装置
音声蓄積部
14
21
表示装置
表示形式生成部
特徴量情報比較部
17
16
22
検索キー音声信号入力装置
検索キー音声時系列特徴量線形圧縮部 31
検索キー音声時系列特徴量抽出部 18
検索キー音声時系列特徴量ベクトル抽出フェーズ P4
音声検索フェーズ P3

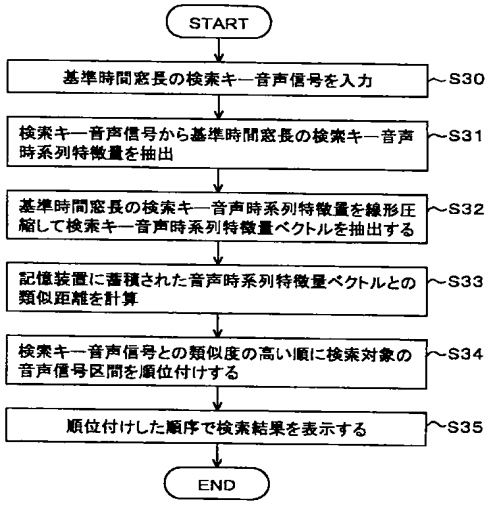
【図８】


START
検索対象の音声信号が入力される S20
音声信号から音声時系列特徴量を抽出 S21
基準時間窓を少しずつずらして部分音声時系列特徴量を切り出す S22
複数種類の長さの時間窓でも部分音声時系列特徴量を切り出す S23
複数種類の長さの部分音声時系列特徴量を線形伸縮させて基準時間窓の長さにする S24
基準時間窓長の部分音声時系列特徴量を線形圧縮して音声時系列特徴量ベクトルを抽出する S25
インデックスを構築し，記憶装置に蓄積 S26
END

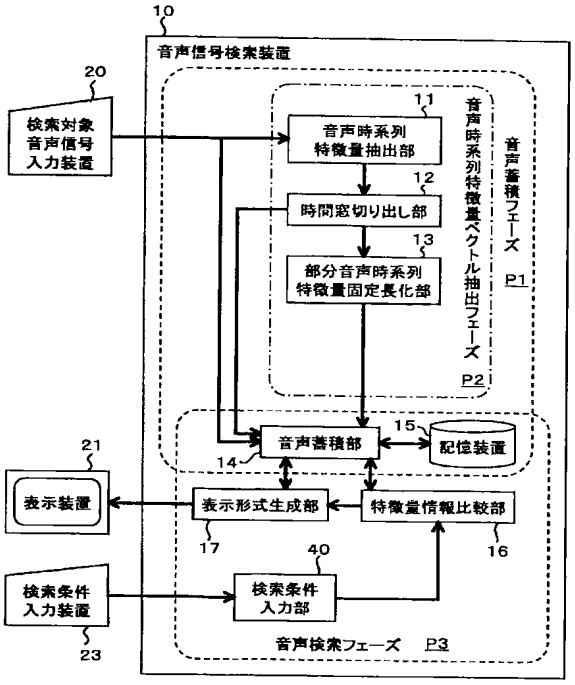
【図９】


時間方向
基準時間窓長の音声時系列特徴量
線形圧縮
音声時系列特徴量ベクトル

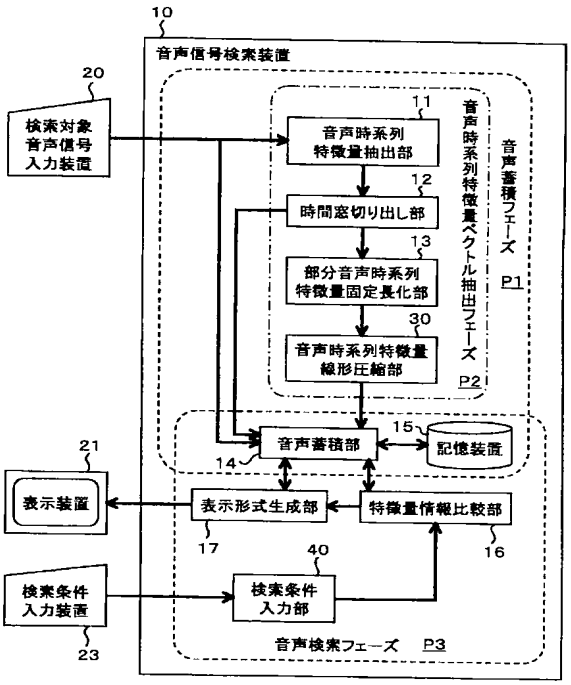
【図１０】


START
基準時間窓長の検索キー音声信号を入力 S30
検索キー音声信号から基準時間窓長の検索キー音声時系列特徴量を抽出 S31
基準時間窓長の検索キー音声時系列特徴量を線形圧縮して検索キー音声時系列特徴量ベクトルを抽出する S32
記憶装置に蓄積された音声時系列特徴量ベクトルとの類似距離を計算 S33
検索キー音声信号との類似度の高い順に検索対象の音声信号区間を順位付けする S34
順位付けした順序で検索結果を表示する S35
END

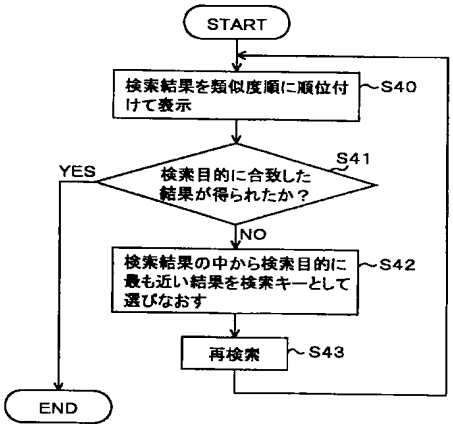
【図１１】



【図１２】



【図１３】



【図１４】

第1の実験の結果

	平均探索長の平均	
歌名	線形伸縮なし	本発明
歌A（LOVE LOVE LOVE）	×	1
歌B（青いイナズマ）	1	1
歌C（DEPARTURES）	×	1．5
歌D（これが私の生きる道）	1．75	1
歌E（チェリー）	1	1
歌F（島唄）	×	4．25
歌G（津軽海峡冬景色）	4．25	2．5
歌H（硝子の少年時代）	1	1
歌I（flower）	4	2
歌J（大切なあなた）	×	1
歌K（今すぐKissMe）	×	4
歌L（ペッパー警部）	1．25	1

【図１５】

第2の実験の結果

歌名	平均探索長の平均	
	従来方法 非線形伸縮	本発明 線形伸縮
歌A（LOVE LOVE LOVE）	1	1
歌B（青いイナズマ）	1	1
歌C（DEPARTURES）	1	1．5
歌D（これが私の生きる道）	1	1
歌E（チェリー）	×	1
歌F（鳥唄）	×	4．25
歌G（津軽海峡冬景色）	1	2．5
歌H（硝子の少年時代）	1．75	1
歌I（flower）	×	2
歌J（大切なあなた）	2．5	1
歌K（今すぐKissMe）	×	4
歌L（ペッパー警部）	×	1

フロントページの続き

(51)Int.Cl.[7]　　識別記号　　Ｆ Ｉ　　ﾃｰﾏｺｰﾄﾞ(参考)
Ｇ１０Ｌ　15/10

(72)発明者　片岡　良治
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

Ｆターム(参考)　5B075 ND14 PP07 PP12 QM05
5D015 DD03 HH01 HH04